Rejection of Claims Under 35 U.S.C. 112, second paragraph

Claims 1-25 have been rejected under 35 U.S.C. 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter of the invention.

The Examiner maintains that claim 1 is incomplete for omitting essential steps and is of the view that at least one step of washing the filter is essential for the method of the invention. The Applicants, however, disagree, as the method allows detection of the amyloid-like fibrils or protein aggregates without an additional step of washing. The step of washing the filter is merely an optional step of the method and thus represents a preferred embodiment. Said preferred embodiment is disclosed in the description as originally filed on page 5, second paragraph. Moreover, dependent claim 8 is directed to a method further comprising a step of washing prior to the step of detection (b). Therefore, it would have been clear to one of skill in the art that a washing step is optional.

The Examiner also maintains that claim 1 is incomplete for omitting a label or a tag or a signal producing system to enable detection of the amyloid-like fibrils or protein aggregates. The present invention relates to a method of detecting the presence of amyloid-like fibrils or protein aggregates in samples. By said method, all suitable ways of detection of the fibrils or aggregates, that would be known to one of skill in the art, are comprised. Accordingly, the detection of the amyloid-like fibrils or protein aggregates with antibodies, peptides or polypeptides, enzymes or fragments or derivatives thereof, or chemical reagents represent a preferred embodiment of the present invention. Dependent claims 10 and 11 are directed to such methods characterized by particularly specified ways of detection of the presence of the fibrils or aggregates. However, these are not essential to the practice of the invention as recited in claim 1 as the size of the amyloid-like fibrils or protein aggregates may allow detection with the naked eye or microscopy. Therefore, the method of claim 1 is not incomplete as it would be clear to a person of skill in the art, that a specific label, tag or signal producing system is not necessary to enable detection of the fibrils or aggregates.

According to the Examiner, claim 2 is ambiguous in reciting "said amyloid-like fibrils or protein aggregates are indicative of a disease" as it is not specifically defined how these fibrils or aggregates are detected so as to be indicative of a disease. The Applicants respectfully disagree. The specification teaches that the detection of the recited fibrils or protein aggregates is indicative for a designated group of diseases. According to a preferred embodiment, said

diseases are human diseases, such as in claim 3. In particular, the correlation of the presence of said fibrils or aggregates with diseases is described in the specification of the present invention on pages 1 and 2 for Alzheimer's disease and Huntington's disease. Additional steps are not required. The identification of the fibrils or protein aggregates is correlative with the disease state. Such correlations have also been suggested in the art, e.g., by Kelly ("Alternative Confirmations of Amyloidogenic Proteins Govern Their Behaviour," *Curr. Opin. Struct. Biol.*, (1996) 6:11-70).

The Examiner also maintains that claim 4 is ambiguous in reciting "said disease is associated with a polyglutamine expansion" and that it is unclear what is encompassed by the term "associated." Claim 4 is directed to a method of detecting the presence of amyloid-like fibrils or protein aggregates which are indicative of disease and wherein said disease is characterized by a significant polyglutamine expansion. The latter feature is recited in claim 4 by the term "wherein said disease is associated with a polyglutamine expansion." The term "associated" is a well known term. If fact, the objected term is used by the Examiner to describe the teachings of Trottier et al., which teaches that polyglutamine expansion can be found in proteins which cause Huntington's disease and other disorders. Thus, the meaning of the term "associated" is unambiguous and clear to a person skilled in the art in the context of claim 4.

The Examiner has rejected claim 5 for being indefinite in reciting "BSE". Claim 5 has, therefore, been amended to clarify the acronym BSE by defining the acronym in the claim.

The Examiner also maintains that claims 6 and 7 are vague and confusing as the same term "material" is used and accordingly claims 6 and 7 are believed to lack clear antecedent support in reciting "material." Claim 1 was amended to recite "material of a sample". The amendment is sufficient to clarify the difference between the material of the sample and the material of claim 6 and 7.

Claims 8 and 9 have been rejected as being vague and confusing in the recitation of "material". Claims 8 and 9 have been amended to clarify that the material is in fact material of the sample.

The Examiner has rejected claim 9 as being confusing because the material of the sample is simultaneously with or subsequent to step (a), sucked through said filter. Step (a) comprises contacting said filter with the material of the sample. Claim 9 is directed to the method further comprising sucking the detergent- or urea-soluble material of the sample through said filter. It is apparent to one of skill in the art that the detergent- or urea-soluble material of the sample can be

sucked through the filter upon simultaneous contact with the filter or subsequent to said contact. Claim 9 is clear on its face to a person of skill in the art.

Claim 10 has been rejected by the Examiner as being indefinite for the recitation of "(poly)peptide". Applicants have amended the claim to clarify that the (poly)peptide is a peptide or polypeptide. The Examiner has further rejected claim 10 as being indefinite in reciting a "fragment or derivative thereof". Claim 10 recites "a step of detection which can be effected by an antibody, or peptide or polypeptide, preferably a tag or an enzyme, or a fragment or derivative thereof or a chemical reagent." It would be apparent to one of skill in the art that such fragment or derivative would comprise any portion of said antibody or peptide or polypeptide that would still effect detection upon binding the fibrils or aggregates. Additionally, on page 6, paragraph 5, the specification provides a description of a fragment being a fragment that retains the function of the peptide or polypeptide.

Claim 12 has been rejected for lacking clear antecedent support for the recitation of "said material" and for a confusing dependence from claim 3, as claim 3 is a method for human disease. Claim 12 has been amended to clarify that the material is the material of the sample. Claim 12 has also been amended to remove its dependency from claim 3. This change clarifies that the material of the sample is derived from bacteria, yeast, plants, humans, etc.

Claim 13 was rejected by the Examiner for the recitation of "(poly)peptide". Claim 13 was also rejected as being indefinite for the recitation of "and/or". Claim 13 has been amended to clarify that the (poly)peptide is in fact a peptide or polypeptide. Claim 13 has further been amended to indicate that the fusion protein comprises a peptide or polypeptide that enhances the solubility or prevents aggregation of said fusion protein. The term "or" in this instance encompasses the possibility that the peptide or polypeptide can both enhance solubility and prevent aggregation of said fusion protein as well as doing only one or the other.

Claim 13 has also been objected to by the Examiner. According to the Examiner step (a') is indefinite when reciting incubating as no specific requirements are recited in the claim. Step (a') of claim 13 clearly describes the step of incubation for the person skilled in the art. Those of ordinary skill in the art are familiar with the meaning of the term "incubation." It is a commonly used procedure in the biological sciences. Furthermore, page 7, first paragraph of the specification, teaches that suitable conditions of incubation may be determined by the person skilled in the art according to conventional procedures.

The Examiner has further objected to claim 13 as step (a") is vague and indefinite in reciting a "compound that induces cleavage" and a "suspected inhibitor" in step (a'). Step (a') describes a fusion protein which comprises a peptide or polypeptide that may enhance solubility or prevent aggregation of the fusion protein or both and also comprises an amyloidogenic peptide or polypeptide which has the ability to self assemble into fibrils or protein aggregates when released from the fusion protein. It is clear to one of skill in the art that releasing the amyloidogenic peptide or polypeptide from the other component of the fusion protein can be effected by a cleavable site that separates these components. Further, it is clear that said fusion protein is intended to be incubated in the presence of a suspected inhibitor of fibril or aggregate formation. Step (a'') provides for the further incubation with a compound that induces cleavage at the cleavage site. It would be clear to one of skill in the art that the two terms are meant to describe different compounds as recited in the claim.

The Examiner has rejected claim 14 as being ambiguous in reciting "site cleavable by intein self cleavage". The meaning and function of said self-processing protein is clear to one of skill in the art. In this context, the Examiner is referred to a publication of Perler et al. (*Nucleic Acids Research,* 1997, 25(6), 1087-1093) (copy attached as Appendix 1) which describes inteins as protein splicing elements. In light of the teachings of the specification and knowledge in the art, the term would be clear to one of skill in the art.

Claim 15 has been rejected by the Examiner as being confusing in reciting the inhibitor of said compound that induces cleavage. It is clear to one of skill in the art that the inhibitor of said compound is an inhibitor of the compound that induces cleavage of the fusion protein.

Claim 16 has been rejected by the Examiner as being indefinite for the recitation of "(poly)peptide". Applicants have amended the claim to clarify that the (poly)peptide is a peptide or polypeptide.

Claim 18 has been rejected by the Examiner for lacking clear antecedent support in reciting "said material". Claim 18 has been amended to clarify that material means material of the sample.

Claim 20 has been rejected by the Examiner as being indefinite for the recitation of "SDS". Claim 20 has been amended to clarify that the acronym SDS stands for Sodium Dodecyl Sulphate. Claim 20 has also been rejected for improperly reciting the trademark "Triton X-100" and has been amended to explicitly indicate it is a trademark.

The Examiner has rejected claims 21-23 and 25 analogously to the rejection of claim 15 in reference to the lack of clarity with the use of the term "inhibitor". Claim 21 has been amended to clarify that the inhibitor is the inhibitor of amyloid-like fibril or protein aggregate formation. Claim 25 was not amended because the claim clearly defines the inhibitor as one that inhibits the compound that induces cleavage defined in step (iii).

Claim 23 was rejected by the Examiner as being in improper form. As claim 23 has been amended and is no longer a multiple dependent claim, the rejection is overcome. Further, claim 23 has been amended, in response to the Examiner's rejection of the recitation of "and/or", to recite a pharmaceutically acceptable carrier or diluent. The amended claim encompasses a composition that includes both a pharmaceutically acceptable carrier and diluent, or either component alone.

Claims 24 and 25 were rejected by the Examiner for improper antecedent bases. Both claims have been amended accordingly.

Rejection of Claims 1-3, 5-12 and 18-19 Under 35 U.S.C. 102(b)

Claims 1-3, 5-12 and 18-19 are rejected under 35 U.S.C. 102(b) as being anticipated by Tateishi et al. The Examiner maintains that Tateishi et al. describes infectious CJD aggregates which are retained by a filter membrane in the absence of detergent.

The Tateishi reference does not anticipate the claimed invention because it is missing at least one element. In the presence of the detergent, Sarkosyl, Tateishi recites that aggregates "were separated into small pieces" (page 361, right column, line 9). Thus, said aggregates were diluted into smaller pieces and no longer retained by the filter membrane but were contained in the corresponding filtrate. The significant increase in the infection ability in the aggregates separated by the use of detergents is shown in Table 1 (page 360). Accordingly, aggregates described in Tateishi et al. were, in fact, soluble by detergent. Claim 1 is a method of detecting the presence of detergent-or urea-insoluble amyloid-like fibrils or protein aggregates. Tateishi is not detecting the presence of insoluble fibrils or proteins.

Rejection of Claims 4 and 17 Under 35 U.S.C. 103(a)

The Examiner has rejected claims 4 and 17 under 35 U.S.C. 103(a) as being unpatentable over Tateishi et al. in view of Trottier et al. and Stott et al. The Examiner maintains that Tateishi et al. as discussed previously combined with Trottier et al. and Stott et al. would render the

neurodegenerative diseases. The Examiner also maintains that treatment of such diseases with compositions comprising inhibitors are further disclosed. Therefore, the Examiner concludes that one of skill in the art would have been motivated to inhibit formation of amyloid fibrils using inhibitors in the form of pharmaceutical compositions. Claims 15 and 21-25 were not obvious for at least the same reasons as discussed above. The methods of the invention are directed to an *in vitro* method comprising contacting a filter with material of a sample suspected to contain fibrils or aggregates that are detergent- or urea-insoluble. Tateishi et al., as argued previously, does not provide for a method of detecting insoluble fibrils or proteins, and the combination of references does not provide additional guidance that would make obvious the rejected claims.

SUMMARY

It is believed that all of the pending claims are now allowable. If the Examiner has any questions or comments, she is encouraged to contact Applicants' representative at the number listed below.

Respectfully submitted,

By: Helen Lockhart

Helen C. Lockhart, Reg. No. 39,248
Wolf, Greenfield & Sacks, P.C.
600 Atlantic Avenue
Boston, MA 02210-2211
Tel. (617) 720-3500

Docket No. V00179/70001(HCL)
Date: June 4, 2002
X06/04/02

MARKED-UP CLAIMS

1. (Amended) A method for detecting the presence of detergent- or urea-insoluble amyloid-like fibrils or protein aggregates on a filter comprising the following steps:

(a) contacting said filter with material of a sample suspected to comprise said fibrils or aggregates; and

(b) detecting whether said fibrils or aggregates are retained on said filter.

5. (Amended) The method of any one of claims 2 to [4] 3 wherein said disease is Huntington's disease, spinal and bulbar muscular atrophy, dentarorubral pallidoluysian atrophy, spinocerebellar ataxia type-1, -2, -3, -6 or –7, Alzheimer disease, [BSE] bovine spongiform encephalopathy (BSE), primary systemic amyloidosis, secondary systemic amyloidosis, senile systemic amyloidosis, familial amyloid polyneuropathy I, hereditary cerebral amyloid angiopathy, hemodialysis-related amyloidosis, familial amyloid polyneuropathy III, Finnish hereditary systemic amyloidosis, type II diabetes, medullary carcinoma of the thyroid, spongiform encephalopathies: Kuru, Gerstmann-Sträussler-Scheinker syndrome (GSS), familial insomnia, scrapie, atrial amyloidosis, hereditary non-neuropathic systemic amyloidosis, injection-localized amyloidosis, hereditary renal amyloidosis, or Parkinson's disease.

6. (Amended) The method of any one of claims 1 to [5] 3 wherein said filter is comprised of material with low protein adsorption.

8. (Amended) The method of any one of claims [1 to 7] 1 to 3 and 7 wherein, prior to step (b), the following step is carried out: (b') washing said filter so as to remove detergent- or urea-soluble material of the sample.

9. (Amended) The method of any one of claims [1 to 8] 1 to 3 and 7 wherein detergent- or urea-soluble material of the sample is simultaneously with or subsequent to step (a), sucked through said filter.

10. (Amended) The method of any one of claims [1 to 9] 1 to 3 and 7 wherein detection in step (b) is effected by an antibody, or [(poly)peptide] peptide or polypeptide, preferably a tag

or an enzyme, or a fragment or derivative thereof or a chemical reagent that specifically binds to said fibrils or aggregates.

11. (Amended) The method of any one of claims [1 to 9] 1 to 3 and 7 wherein detection in step (b) is effected by electron microscopy, electron scanning microscopy, fluorescence or chemiluminescence.

12. (Amended) The method of any one of claims [1 to 11] 1, 2, and 7 wherein said material of the sample is derived from tissues or cells of bacteria, yeast, fungi, plants, insects, animals, preferably mammals, humans, from a transgenic animal or a transgenic plant.

13. (Amended) The method of any one of claims [1 to 11] 1 to 3 and 7 further comprising the following steps prior to step (a):

(a') incubating a fusion protein comprising a [(poly)peptide] peptide or polypeptide that enhances solubility [and/or] or prevents aggregation of said fusion protein, an amyloidogenic [(poly)peptide] peptide or polypeptide that has the ability to self-assemble into amyloid-like fibrils or protein aggregates when released from said fusion protein and a cleavable site that separates the above-mentioned components of the fusion protein in the presence of a suspected inhibitor of amyloid-like fibril or protein aggregate formation; and

(a'') simultaneously with or after step (a'), further incubating with a compound that induces cleavage at said cleavage site.

15. (Amended) The method of claim [13 or] 14 further comprising, prior to step (b) and after step (a''):

(a''') incubation with an inhibitor of said compound that induces cleavage.

16. (Amended) The method of [any one of claims 13 to 15] claim 14 wherein said amyloidogenic [(poly)peptide] peptide or polypeptide comprises a polyglutamine expansion.

17. (Amended) The method of [any one of claims 4 to 16] claim 4 wherein said polyglutamine expansion comprises at least 35, preferably at least 41, more preferably at least 48 and most preferably at least 51 glutamines.

18. (Amended) The method of any one of claims [1 to 17] 1 to 3, and 7 wherein said contacting is effected by dotting, spotting or pipetting said material of the sample onto said filter.

19. (Amended) The method of any one of claims [1 to 18] 1 to 3, and 7 wherein said filter is a filter membrane.

20. (Amended) The method of any one of claims [1 to 19] 1 to 3, and 7 wherein said detergent is [SDS] Sodium Dodecyl Sulphate (SDS) or [Triton X-100] TRITON X-100™.

21. (Amended) An inhibitor of amyloid-like fibril or protein aggregate formation identified by the method of [any one of claims 13 to 19] claim 14.

23. (Amended) A pharmaceutical composition comprising the inhibitor of claim [21 to] 22 and a pharmaceutically acceptable carrier [and/or] or diluent.

24. (Amended) A diagnostic composition comprising

(i) [a] the fusion protein [as defined in] of any one of the preceding claims.

25. (Amended) The diagnostic composition of claim 24 further comprising

(ii) [a] the filter [as defined in] of any one of the preceding claims optionally or preferably contained in a microtiter plate; and optionally

(iii) [a] the compound that induces cleavage [as defined in] of any one of the preceding claims; and optionally;

(iv) an inhibitor of said compound of [(c)] (iii); and optionally

(v) suitable buffer solutions.

MARKED-UP SPECIFICATION

On page 5, last sentence of the fourth paragraph, delete "Triton X-100" and replace with --TRITON X-100™--.

Step (b') may be repeated one or several times. The person skilled in the art is in a position to determine appropriate washing conditions without further ado. Preferably, the washing buffer comprises 0.1-2% SDS, 4-8M urea, and 0.1-2% [Triton X-100] TRITON X-100™.

On page 8, last line of the third paragraph, delete "Triton X-100" and replace with --TRITON X-100™--.

It is furthermore preferred that the filter is a filter membrane which is optionally or preferably contained in a microtitre plate. Additionally preferred is the use of SDS as detergent or [Triton X-100] TRITON X-100™ for non-β-amyloid aggregates.

On page 25, fourth paragraph, fifth line, delete "Triton X-100" and replace with --TRITON X-100™--.

Cells were washed with buffer A [50 mM sodium phosphate (pH 8), 150 mM NaCl, and 1 mM EDTA]. If necessary, the cell pallet was stored at -70°C. Cells were resuspended in 25 ml buffer A. PMSF and lysozyme (Boehringer Mannheim) were added to 1 mM and 0.5 mg/ml, respectively, and incubated on ice for 45 min. Cells were lysed by sonication (2 x 45 s, 1 min cooling, 200-300 Watt), and [Triton X-100] TRITON X-100™ was added to a final concentration of 0.1% (v/v). The lysate was centrifuged at 30.000 x g for 30 min, and the supernatant was collected.

On page 25, paragraph five, line 4, delete "Triton X-100" and replace with --TRITON X-100™--.

5 ml of a 1:1 slurry of GST-agarose (Sigma), previously equilibrated in buffer A, was added and the mixture was stirred for 30 min. The slurry was poured into a 1.6 cm diameter column, washed once with 40 ml buffer A containing 1 mM PMSF and 0.1% [Triton X-100] TRITON X-100™ and twice with 40 ml buffer A containing 1 mM PMSF. The protein was eluted with 5 x 2 ml buffer A containing 15 mM reduced glutathione (Sigma). Aliquots of the fractions were analyzed by SDS-PAGE and the fractions containing purified GST fusion protein were combined. Finally, the pooled fractions were dialysed

Appendix 1

© 1997 Oxford University Press Nucleic Acids Research, 1997, Vol. 25, No. 6 1087–1093

Compilation and analysis of intein sequences

Francin B. P rler*, Gary J. Ols n[1] and Eric Adam

New England Biolabs Inc., 32 Tozer Road, Beverly, MA 01915, USA and [1]Department of Microbiology, University of Illinois, B301 C&LSL, 601 South Goodwin Avenue, Urbana, IL 61801, USA

Received November 20, 1996; Revised and Accepted January 24, 1997

ABSTRACT

We have compiled a list of all the inteins (protein splicing elements) whose sequences have been published or were available from on-line sequence databases as of September 18, 1996. Analysis of the 36 available intein sequences refines the previously described intein motifs and reveals the presence of another intein motif, Block H. Furthermore, analysis of the new inteins reshapes our view of the conserved splice junction residues, since three inteins lack the intein penultimate His seen in prior examples. Comparison of intein sequences suggests that, in general, (i) inteins present in the same location within extein homologs from different organisms are very closely related to each other in paired sequence comparison r phylogenetic analysis and we suggest that they should be considered intein alleles; (ii) multiple inteins present in the same gene are no more similar to each ther than to inteins present in different genes; (iii) phylogenetic analysis indicates that inteins are so divergent that trees with statistically significant branches cannot be generated except for intein alleles.

INTRODUCTION

Protein splicing is defined as removal of an *int*ernal prot*ein* segment (*intein*) from a precursor protein and ligation of the *ex*ternal pro*tein* segments (*exteins*) to form a native peptide bond (1). Extein ligation differentiates protein splicing from other forms of self-proteolysis, such as cleavage of glycosylasparaginase (2) or the hedgehog protein (3). Protein splicing elements were first described in 1990 as in-frame insertions in the *Saccharomyces cerevisiae* VMA gene that were unrelated to the sequence of homologous ATPases (4,5). Moreover, the mature VMA protein had an electrophoretic mobility that was similar to the homolog lacking the intein and not to the predicted size of the VMA gene. A second protein with the predicted size of the intein was also detected. Most inteins contain the dodecapeptide motifs characteristic of homing endonucleases (which were first discovered in mobile self-splicing introns) and several inteins have demonstrated endonuclease activity (6–10). Intein genes that encode active homing endonucleases are potential mobile genetic elements (6,11,12).

Although several inteins were identified experimentally (inteins 1–3, 6, 7, 11, 12, 14, 15 and 18 in Table 1) (4,5,10,13–16; Cole,S., personal communication; Liu,P.X.-Q., personal communication), most of the recently described inteins were predicted from DNA sequences (9,15,17–20). This latter class of inteins is termed theoretical in Table 1, since spliced products have not been experimentally observed. A combination of four criteria have been used to identify protein splicing elements in newly sequenced genes (9,15,17,18): (i) an in-frame insertion in a gene that has a previously sequenced homolog lacking the insertion; (ii) the presence of intein Blocks C and E (Table 2), which are also found in homing endonucleases, where they are called dodecapeptide motifs, DOD motifs, P1 and P2 motifs and LAGLI-DADG motifs (8,9); (iii) the presence of several other conserved intein motifs (Table 2; 9); (iv) the presence of four conserved splice junction residues (Ser, Thr or Cys at the intein N-terminus, the dipeptide His–Asn at the intein C-terminus and Ser, Thr or Cys following the downstream splice site) (1,9,21–24). The last three criteria help differentiate true inteins from in-frame inserts that result from interspecies sequence variability or other types of insertion sequences. As discussed below, these criteria have been refined as more inteins have been discovered.

ANALYTICAL METHODS

Alleles with >41% identity to the prototype intein first identified in that location, as determined using the default parameters of the BESTFIT pairwise comparison program (25), were not included in the multiple sequence analysis, since they would bias the search for conserved motifs and the calculation of their significance. This percent identity was chosen because it is just above the highest identity among the poorly related intein alleles (see below). The *Mka* gyrA, *Mfl* gyrA, *Mgo* gyrA, *Mxe* gyrA, *Psp* pol-1, *Psp* pol-3 and *Mja* pol-2 inteins were not included while building the alignment nor were they included in the block calculations.

Conserved motifs were detected and evaluated with the MACAW 2.0.5 program (26). MACAW does not allow gaps in the aligned sequence blocks. Briefly, the Gibbs sampling method (27) was used for identifying the sequence blocks and the block boundaries were readjusted to maximize the motif score (minimizing the *p* value) using the BLOSUM62 comparison matrix (28). The MACAW program calculates the chance probability for the appearance of an alignment score by a statistical formula

* To whom correspondence should be addressed. Tel: +1 508 927 5054; Fax: +1 508 921 1350; Email: perler@neb.com

using an extreme value distribution model of alignment scores (*p* value) (26). All the final block calculations resulted in *p* values $<10^{-20}$ (i.e. the calculation limit of the MACAW 2.0.5 program on the Power Macintosh 8100/80) when the whole length of the 29 most diverse intein sequences (as defined above) was taken as the search sequence space.

The least squares distance phylogenetic tree was inferred using programs in version 3.5 of the PHYLIP package (29). The number of amino acid replacements per sequence position separating each pair of sequences was estimated using the PAM option of the PROTDIST program. The sampling variance of the distance values was estimated from 100 bootstrap resamplings of the sequence data using the SEQBOOT and PROTDIST programs. The phylogenetic tree that best fits (by a least squares criterion) these sequence-to-sequence distances was found with the FITCH program, using the subreplicates option to weight each pairwise distance by one over its estimated variance (30). Global rearrangements and multiple taxon addition orders were used to find an optimal tree. Because of possible errors in Block E of the *Psp* pol-3 intein sequence, three positions were replaced by unidentified residues (Xs) in the phylogenetic analysis, yielding FLEGXXXGDG.

THE CATALOG

The information summarized in Table 1 comprises all intein sequences that to our knowledge have been published or were available from public databases [NCBI sequence libraries or The Institute for Genomic Research (TIGR) Web page, http://www.tigr.org] as of September 18, 1996. Inteins whose sequences were not available have not been included in this list. Updates to this catalog can be obtained via Email from perler@neb.com and new inteins can be registered at this same address. The registry will also be accessible in the near future on the New England Biolabs Web site (http://www.neb.com). The REBASE database (http://www.neb.com/rebase) also collects information about inteins, with emphasis on endonuclease activity (31).

According to intein nomenclature conventions (1), the intein names listed in Table 1 include organism and extein gene designations as well as a numerical suffix when more than one intein is present in the same extein gene in the same organism (as in the case of the *Tli* and *Mja* pol inteins, *Mja* RNR inteins and *Mja* RFC inteins). DNA polymerase inteins from various *Pyrococcus* isolates (*Psp* pol inteins 1–3) were numbered in order of entry into the intein registry and are not present in the same organism (Table 1). The *Mle* recA intein and the *Mtu* recA intein are located at different positions in recA (after G205 or K251 respectively). There are also many examples of inteins present in the same location in homologous extein genes from different organisms (dnaB, VMA, pol and gyrA). If endonuclease activity has been demonstrated, the intein is also given an endonuclease designation following the restriction enzyme nomenclature convention with the addition of the prefix PI-. To date, four inteins have demonstrated endonuclease activity: PI-*Sce*I (*Sce* VMA intein), PI-*Tli*I (*Tli* pol intein-2), PI-*Tli*II (*Tli* pol intein-1) and PI-*Psp*I (*Psp* pol intein-1) (7,10,32; Perler,F.B., unpublished data).

Except for the *Sce* VMA intein, the *Tli* pol-2 intein and the *Psp* pol-1 intein, for which N-terminal amino acid sequences have been determined (10,24,33), the size and splice junction residues listed in Table 1 have been deduced using the criteria listed above for theoretical inteins (4,5,9,10,13–20,34). Exact intein boundaries are usually obvious after comparison with inteinless homologs, especially since many inteins are present in conserved motifs in extein genes, such as DNA polymerases and gyrases (15,35). The TIGR Web site alignments were used to determine *M.jannaschii* intein boundaries, except for the *Mja* hyp-1 intein, where the *Bacillus subtilis* YqkH protein (GenBank accession no. D84432) provided a better extein match than the *B.subtilis* YqxK protein (36). Extein sequences flanking each *M.jannaschii* intein were not always similar to the sequence of the inteinless homolog. In these cases, the intein boundaries were deduced by comparison with conserved sequences in Blocks A and G (see below and Table 2) and are marked with an asterisk in Table 1. However, because of the high degree of conservation of the intein junctions and other residues in Blocks A and G, the presence of an asterisk does not imply reduced confidence in junction assignment.

Inteins have been found in all three domains of life (Table 1): (i) inteins 1–2 are in eucaryal nuclear genes (*Sce* VMA and *Ctr* VMA) and inteins 3–4 are in eucaryal chloroplast genes (*Ceu* clpP and *Ppu* dnaB); (ii) inteins 5–13 are from eubacteria (*Mycobacterium* and *Synechocystis* spp.); (iii) inteins 14–36 are from thermophilic Archaea (*Thermococcus litoralis*, *Pyrococcus* isolates and *Methanococcus jannaschii*). Inteins are found in the same types of organisms and chromosomal locations as mobile introns (37). The large number of inteins reported in *Mycobacterium leprae* and *M.jannaschii* are due, in part, to genome sequencing projects. However, only one intein has been found in the genomes of *Synechocystis* spp. (38) and *S.cerevisiae* and no inteins have been detected in *Haemophilus influenzae* Rd (39), *Mycoplasma genitalium* (40) and other viral or phage genome sequences present in GenBank as of September 18, 1996. Whether the 18 inteins in 14 different *M.jannaschii* genes (17) reflect an abundance of inteins in this particular species or in Archaea in general awaits a complete analysis of more small genomes. For now we note that extensive sequencing of archaeal RNA polymerase genes (41–45) and DNA polymerase genes (35) suggests that these inteins are not widely distributed in Archaea.

Although many inteins are located in enzymes that interact with nucleic acids, several inteins are located in metabolic enzymes, such as phosphoenolpyruvate synthase, anaerobic ribonucleoside triphosphate reductase, UDP-glucose dehydrogenase, ClpP protease/chaperone, vacuolar ATPase proton pump (VMA) and glutamine-fructose 6-phosphate transaminase (Table 1).

The inteins listed in Table 1 range in size from 335 to 548 amino acids, except for the *Ppu* dnaB intein (150 amino acids) and the *Mxe* gyrA intein (198 amino acids). The central domain present in other inteins is missing in the *Mxe* gyrA (GenBank accession no. U67876) and *Ppu* dnaB (18) inteins (Table 2). These small inteins may have lost those residues required for endonuclease activity and may thus represent minimal inteins. Alternatively, they may represent an intein remnant that is no longer capable of splicing.

We suggest that inteins present in the same position in an extein homolog from different organisms should be designated *intein alleles*. *Psp* pol intein-1 and *Tli* pol intein-1 alleles have the same endonuclease specificities (Perler,F.B., unpublished data). Pairwise amino acid sequence comparisons indicate that the 11 inteins present in identical locations in DNA polymerase or gyrA genes are more similar to their alleles than to any other intein (at least

Table 1.

No.	Intein Name	Extein Name	Organism	Allele	Type	N-term	C-term	Size	Loc	Acc No.	Ref
Eucarya											
1	† *Sce* VMA	Vacuolar ATPase, subunit	*S. cerevisiae*		Exp	C	HN/C	454	G283	M21609	4-7
2	*Ctr* VMA	Vacuolar ATPase, subunit	*C. tropicalis*	*Sce* VMA	Exp	C	HN/C	471	G283	M64984	16
3	*Ceu* clpP	clpP	*C. eugametos*		Exp	C	GN/S	456	E447	L29402	∞,20
4	*Ppu* dnaB	DnaB helicase	*P. purpurea*		Theor	C	HN/S	150	G361	U38804	19
Eubacteria											
5	*Ssp* dnaB	DnaB helicase	*Synechocystis*	*Ppu* dnaB	Theor	C	HN/S	429	G361	D64003	18
6	*Mtu* recA	RecA	*M. tuberculosis*		Exp	C	HN/C	440	K251	X58485	13,34
7	*Mle* recA	RecA	*M. leprae*		Exp	C	HN/S	365	G205	X73822	14
8	*Mle* pps1	Pps1	*M. leprae*		Theor	C	HN/S	386	G201	U00013	9
9	*Mle* gyrA	GyraseA	*M. leprae*		Thero	C	HN/T	420	Y130	Z68206	§,15
10	*Mka* gyrA	GyraseA	*M. kansasii*	*Mle* gyrA	Theor	C	HN/T	420	Y130	Z68207	15
11	*Mfl* gyrA	GyraseA	*M. flavescens*	*Mle* gyrA	Exp	C	HN/T	421	Y130	Z68209	§,15
12	*Mgo* gyrA	GyraseA	*M. gordonae*	*Mle* gyrA	Exp	C	HN/T	420	Y130	Z68208	§,15
13	*Mxe* gyrA	GyraseA	*M. xenopi*	*Mle* gyrA	Theor	C	HN/T	198	Y130	U67876	47
Archaea											
14	† *Tli* pol-1	DNA polymerase	*T. litoralis*		Exp	S	HN/S	538	N494	M74198	10
15	† *Psp* pol-1	DNA polymerase	*Psp* GB-D	*Tli* pol-1	Exp	S	HN/S	537	N492	U00707	33
16	*Psp* pol-3	DNA polymerase	*Psp* KOD	*Tli* pol-1	Theor	S	HN/S	536	N851	D29671	60
17	*Mja* pol-2	DNA polymerase	*M. jannaschii*	*Tli* pol-1	Theor	S	HN/S	476	N882	U67532	17
18	† *Tli* pol-2	DNA polymerase	*T. litoralis*		Exp	S	HN/T	390	D1081	M74198	10
19	*Psp* pol-2	DNA polymerase	*Psp* KOD		Theor	C	HN/S	360	R406	D29671	60
20	*Mja* pol-1	DNA polymerase	*M. jannaschii*	*Psp* pol-2	Theor	C	HN/S	369	R425	U67532	17
21	*Mja* hyp-1	Hypothetical protein-1	*M. jannaschii*		Theor	C	HN/C	392	H128	U67462	17
22	*Mja* hyp-2	Hypothetical protein-2	*M. jannaschii*		Theor	C	HN/C	488	N97	U67515	17
23	*Mja* IF-2	Translation initiation factor	*M. jannaschii*		Theor	C	HN/T	546	K30	U67481	17
24	*Mja* TFIIB	Transcription factor IIB	*M. jannaschii*		Theor	*S	HN/T	335	Y99	U67522	17
25	*Mja* PEP Syn	PEP synthase	*M. jannaschii*		Theor	C	FN/C	412	T410	U67503	17
26	*Mja* RNR-1	Anaerobic rNTP reductase	*M. jannaschii*		Theor	*S	*HN/T	453	Q337	U67527	17
27	*Mja* RNR-2	Anaerobic rNTP reductase	*M. jannaschii*		Theor	*S	*HN/T	533	S1058	U67527	17
28	*Mja* Rpol A''	RNA polymerase subunit A''	*M. jannaschii*		Theor	S	HN/T	471	M75	U67547	17
29	*Mja* Rpol A'	RNA polymerase subunit A'	*M. jannaschii*		Theor	C	GN/C	452	V463	U67547	17
30	*Mja* UDP GD	UDP-glucose dehydrogenase	*M. jannaschii*		Theor	*C	*HN/C	454	S260	U67548	17
31	*Mja* Helicase	Helicase	*M. jannaschii*		Theor	C	HN/S	501	L337	U67555	17
32	*Mja* GF-6P	GF-6P transaminase	*M. jannaschii*		Theor	C	HN/S	499	H74	U67582	17
33	*Mja* r-gyr	Reverse gyrase	*M. jannaschii*		Theor	C	HN/C	494	L866	U67592	17
34	*Mja* RFC-1	Replication factor C	*M. jannaschii*		Theor	C	HN/T	548	K53	U67583	17
35	*Mja* RFC-2	Replication factor C	*M. jannaschii*		Theor	S	HN/S	436	A626	U67583	17
36	*Mja* RFC-3	Replication factor C	*M. jannaschii*		Theor	C	HN/C	543	S1124	U67583	17

Inteins 1–4 are from eukarya, inteins 5–13 are from eubacteria and inteins 14–36 are from archaea. The *Ceu* clpP intein has also been referred to as IS2 (20). *The exact intein junction was deduced from conserved intein features and not extein similarity. †Endonuclease activity has been demonstrated; however, there are no published activity assays for the other inteins. Allele lists the prototype intein at this same position in a homologous extein gene. N-term and C-term list the residues present at the respective ends of each intein, including the first extein residue following the C-terminal splice junction. Size indicates the number of amino acids in each intein. Loc lists the extein amino acid preceding the intein. The Loc of the *Mxe* gyrA intein was inferred from the other gyrA alleles, since the complete *Mxe* gyrA gene has not been sequenced (GenBank accession no. U67876). §Cole,S., personal communication. ∞Liu,P.X.-Q., personal communication. Other abbreviations: Theor, theoretically derived; Exp, experimentally determined; (/), splice junction; Acc No., accession no.; Ref, reference; pol, DNA polymerase; hyp, hypothetical protein; IF-2, translation initiation factor, FUN12/bIF-2 family; PEP synthase, phosphoenolpyruvate synthase; RNR or anaerobic rNTP reductase, anaerobic ribonucleoside triphosphate reductase; Rpol, RNA polymerase subunit; GF-6P transaminase, glutamine-fructose 6-phosphate transaminase; Replication factor C, replication factor C 37 kDa subunit; *C.tropicalis*, *Candida tropicalis*; *C.eugametos*, *Chlamydomonas eugametos*; *P.purpurea*, *Porphyra purpurea*; *Ssp* or *Synechocystis*, *Synechocystis* spp.; *Psp*, *Pyrococcus* spp.; *M.tuberculosis*, *Mycobacterium tuberculosis*; *M.kansasii*, *Mycobacterium kansasii*; *M.flavescens*, *Mycobacterium flavescens*; *M.gordonae*, *Mycobacterium gordonae*; *M.xenopi*, *Mycobacterium xenopi*.

~60% identity, except for the *Mja* pol-2 intein, which is only 40.4% identical to the *Tli* pol-1 intein). Identity among non-allelic inteins is quite low, generally ranging from 15 to 30%. The VMA inteins are 36.6% identical and branch together in phylogenetic trees (Fig. 1). The only intein alleles that fail to phylogenetically group together are the dnaB alleles (23% identical), possibly because 46 out of 95 residues used in this analysis are absent in the *Ppu* dnaB mini-intein. However, it is difficult to determine whether very dissimilar intein alleles arose from different ancestors or by divergence.

CONSERVED RESIDUES AND THE PROTEIN SPLICING MECHANISM

Protein splicing is so rapid that the precursor protein is rarely observed. The intein plus the first downstream extein residue contain sufficient information for splicing in foreign proteins (13,24,33). However, the exteins may affect splicing rates or efficiency. Using a chimeric intein construct, *in vitro* splicing of a purified precursor was demonstrated (33) and the chemical mechanism of protein splicing was determined (21,33,46–49).

Table 2. Conserved motifs found in inteins

No.	Intein	Block A	Block B	Block C	Block D	Block E	Block H	Block F	Block G
Eucarya									
1.	*Sce* VMA	CFAKGTNVLMADG 13	LLKYTCNRTHELVV 83	LLGLWIGDG 219	VKSIPSFL 307	FLAGLIDSDG 327	TISTSVRDGLVSLARSLGL 359	YGITLSDDSDHQFL 445	HQVVVHNC 455
2.	*Ctr* VMA	CFTRGTQVMMADG 13	LMDFTVSADHKLIL 78	LLGTWAGIG 210	VKSIPQHI 325	LIAGLVDAAG 345	TSFRHVARGLVKIARSLGI 379	YGIYLAKETDHQFL 462	HNALVHNC 472
3.	*Ceu* clpP	CLTSDHTVLTTRG 13	GVDLFVTPNHHMYV 73	PFGLWIANG 151	HKYLPDWV 230	——	STSERFANDVSRLALSAGT 281	PVYCLTGPNHVFTV 445	KAVWTGHS 457
4.	*Ppu* dnaB	CISKFSHIMWSHV 13	EKYLELTSNHKILT 74	——	——	——	——	SVFDFAAMPIPHFI 142	HNIIVHNS 151
Eubacteria									
5.	*Ssp* dnaB	CISGDSLISLAST 13	GKYIRATANHRFLT 77	LLGHLIGDG 122	KKFVPHQV 207	FLRSLWSTDG 227	TSSKKLAKDVQSLLLKLGI 263	EVFDLTVPGPHNFV 421	HDIIVHNS 430
6.	*Mtu* recA	CLAEGTRIFDPVT 13	GAIVHATPDHKVLT 77	LLGYLIGDG 123	EKFIPNMF 201	LLPGLFESDG 223	TTSRQLAHQIHNLLLRFGV 257	RTFDLEVEELHTLV 432	EGVVVHNC 441
7.	*Mle* recA	CHNYSTRVTLADG 13	KSQFAATPNHLIRT 83	VLGSLMGDG 123	——	LQRAVTLGDG 194	FSLKEYLKALTPLVLAIWT 215	HRFDIEVEGNHNYF 357	DGVMVHNS 366
8.	*Mle* ppel	CLTADARIDVRGK 13	GRALKATGNHQYLV 72	LLGLWLGDG 151	TKRLPAWI 225	LIGGLVDADG 245	FASRELLEDVRQLAIGCGL 276	PTYDIQVVGLHNFV 378	HGIVAHNS 387
9.	*Mle* gyrA	CVSGNSLVRLLFG 13	GYEITGTSNHPLLC 79	LFGAFISGG 134	DKLVPDWL 212	FLQALFEGEG 232	TLSERLAADVQQMLLEFGV 266	PVFSLSVDTEDHSF 411	NGFISHNT 421
10.	*Mka* gyrA	cvtgdalvrlpfg 13	gyevtgtanhpllc 79	llgafiseg 134	dtyvpewm 212	flqalfegdg 232	tvekqlaadvqqmllefgv 266	pvyslrvdtddhaf 411	ngfvshnt 421
11.	*Mfl* gyrA	cvtgdalvrlpfg 13	gyevtgtanhpllc 79	hlgafiseg 134	dkavpewl 212	flqalfegdg 232	tregrlakdiqqmllefgv 266	pvyslrvdtddhsf 412	ngfvshnt 422
12.	*Mgo* gyrA	cltgdalvrlpfg 13	gyevtgtsnhpvlc 79	llgafiseg 134	dksvpewl 212	flqalfegdg 232	trsrqlaidvqqmllefgv 266	pvyslrvdsedhaf 411	ngfvshnt 421
13.	*Mxe* gyrA	citgdalvalpeg 13	glrvtgtanhpllc 79	——	——	——	——	pvyslrvdtadhaf 189	ngfvshnt 199
Archaea									
14.	*Fli* pol-1	SILPNEWLPIIKN 13	GRKIKITAGHSLFT 100	LLGYYVSEG 290	NKRIPSVI 365	FLEAYFTGDG 385	TKSKLLAMQLVFLLKSLGI 415	YVYDLSVEDNENFL 528	GLLYAHNS 539
15.	*Psp* pol-1	ailpeewvplikn 13	grkititeghslfv 100	llgyyvseg 289	nkrvpevi 364	flegyfigdg 384	tksellvnqlvlllnslgv 414	yvydlsvdadenfl 527	gflyahns 538
16.	*Psp* pol-3	silpeewlpvlee 13	grrikiteghelfs 100	llgyyvseg 289	nkripefv 364	flegyfigdg∞384	nekralenqlvllinsvgv 413	yvydlsvednenfl 526	glvyahns 537
17.	*Mja* pol-2	ailpdeyltiiee 13	grtikvtrghslfk 100	flgffvtrq 289	kkhipeel 364	——	akdekylnqlmilfnlvgi 413	yvydlsvednenfl 468	nniyahns 477
18.	*Fli* pol-2	SVSGKSKIIIRGN 13	SWYIDVTEDHSLIG 92	LVGLIVGDG 156	RRKIPNFM 234	FLRGLFSADG 254	NIDADFLREVRKLLWTVGI 287	YVYDIEVEHTERFF 382	HNILVHNT 391
19.	*Psp* pol-2	CHPADTKVVVKGK 13	IHGLKCTPNHKLPV 64	LAGILLAEG 126	VKKIMDHI 209	VLRGFFEGDG 226	TKHNVKIKLVSKLLSQLGI 259	KVYDLTLEDTPYYF 352	HGILTHNS 361
20.	*Mja* pol-1	CNPKGTKVVVKGK 13	VHGLKCTPNHKIPL 64	LIGILLAEG 135	——	ILRGFFEGDG 235	THEYDKIKFIASLLDRLGI 268	SVYDLTLEGRPYYF 361	HGILTHNS 370
21.	*Mja* Hyp-1	CVPPDTLLIILENG 13	PHKIILTPEHPVYA 70	——	RSRIPEKI 156	RLVGYFLSEG 173	TTSKILMNQLRLISLRLGF 293	FVYNLEVEDDSSYV 384	VSGTLHNC 393
22.	*Mja* Hyp-2	CLTSNSKILTDDG 13	GKVLSGSKDHPVLT 76	LLGFAFGDG 156	IYKIPEWI 254	FLAGLFGADG 274	ENILEFLMEIKLLLARFDI 317	KLYDVGIVSHNHNF 479	HSIVVHNC 489
23.	*Mja* IF-2	CLMPNHKVLTKYG 13	WHSITTTTPEHPFLT 90	FAGVMFGDG 215	HIKIPQIL 286	FIEGYFDADG 306	SASKEFIEGLSILLLRFEI 337	YVYDLTTETENFIA 539	HGIVVHNT 547
24.	*Mja* TFIIB	SVDYNEPIIIKEN 13	HKKVRVTRSESVFT 97	ILGYLIAEG 147	NKRIPSII 216	FIDGIYMGKD 236	FVSKKLAEDVIFLLLQIKE 258	YAYDLTVPNAENFV 325	GGYVLHNT 336
25.	*Mja* PEP Syn	CIEGDAKILTDRG 13	KDYIKIYPDHKFPV 83	LGGAVLSDG 134	SRKIPSEI 196	LIAGFVDGDG 220	SSHIKKIEGLIVGLYRLGI 260	EVIKIYVKAENRFH 395	TPIVVFNC 413
26.	*Mja* RNR-1	SLGRDELIFIKDG 13	GYSIIVTEDHSLFN 103	FLGLFVAEG 178	HKRSPEYI 247	FLGGLISGDG 267	TTSEQLLGQLHLLLSDLGM 297	YVYDLSVEDNENFI 444	TGILCHNT 454
27.	*Mja* RNR-2	SLFYDEKILIFKN 13	GKKVRVTEGDHSVFT 101	LIGAFLSEG 236	NKKIPSIL 314	LIEGYIDGDG 333	TTSKTLADTLCLALHKLGI 368	YVYDISVEGTHNFI 523	GFICLHNT 534
28.	*Mja* Rpol A''	SLFYEEKILIIKDG 13	GKEITATPYHSFVI 98	FIGIYLAEG 181	TKKIAKFV 254	LIEGYFDGDG 274	SNSKELIDGIAILLARFNI 305	YVYDISVEGLHTFT 462	DGVLTHNT 472
29.	*Mja* Rpol A'	CVDGDTPVLLDGK 13	GRKIIATKDEPFYT 76	——	RKLIPLTY 146	RIVGHVMGDG 165	YSVKVRKKSLCILLKALGC 244	YVYDIYTISETHNF 443	HGFLTGNC 453
30.	*Mja* UDP GD	CFHPDSVLFIDRG 13	GREIKITTKDSPVVI 81	LIGYYLSEG 216	NKKIPPQM 299	FLKGLFRGDG 319	TVSKKMASSLLILLQLLGI 354	YVYSVETENSLLIT 446	YGILISNC 455
31.	*Mja* Helicase	CLNANTKILGKSG 13	GLKIYTTYPEHIFLV 72	FIGYFIGDG 122	WKHIDAFC 197	LIAGLFDSDG 216	SISKKLVEQLQFVLLRFGI 247	YAIDIKLPDDGSHS 489	HGFIVHNS 502
32.	*Mja* GF-6P	CLSPDTYVILPDG 13	FSKLIYTGKHKLFV 72	IIGYIIGDG 215	HKKTPKYV 291	YLRGIFDAEG 311	HTSKCYIKKTQFLLLRFGI 342	YVYDLEVEDYHNFI 491	HLIIHNHS 500
33.	*Mja* r-gyr	CLTPDTYVVLGDG 13	NYKLKATPDECLLV 77	FAGLVLGDG 212	IYSLPEST 281	LIAGYFDTDG 294	SKKKDVLKKIGIYLMSIGI 333	YVYDLSIKGNQNFI 486	HGVISHNC 495
34.	*Mja* RFC-1	CLTGDTKVIVNGS 13	GKKLKVYTTEPLLI 85	WLGYFIGDG 134	KVRIPKKI 211	FLRAYFDCDG 229	TASKEMAEDLVYALLRFGI 258	TIYDLSVPRYHNFI 538	LPYILHNT 549
35.	*Mja* RFC-2	SVSKDTPILVKKD 13	GGYIKLTGNHSIMN 91	MLGLYVAEG 162	NKRIPDKI 251	FLKGLADGDS 271	SKSDHLLIDTVHLARISGI 301	FVYDVSVPSNHNFV 426	VPILLHNS 437
36.	*Mja* RFC-3	CLTGDAKIYLPDE 13	GKEIKATGDHKPLT 76	LLGFIIGDG 232	GYKIPKWI 321	FLEGLFGADG 341	DKTLEPFEEVKKDLKRFEV 384	DVIDIYCHKDPSFI 535	HGFVSHNC 544
	Consensus	Ch Dp hhh G	G h hF B hhh	LhG hhaG	K IP h	L GhFahDG	p S hh b LL hGI	rVYDLpVe Fh	SGhhhHNp
	Sce HO	HLSKDPTILMANG 13	RLALQCTAGHKLSV 89	HLGLWLGDG 223	EKQIPEFM 314	FLAGLIDSDG 334	TVTSSIMDGIVHISRSLGM 370	TVYGLTIEGSKNFL 456	——

Eight conserved intein motifs were identified by multiple sequence analysis (MACAW) of the 29 inteins listed in capital letters, as described in the text. Intein sequences in lower case are highly similar alleles that were not included in the multiple sequence analysis. These motifs are similar to the previously defined intein blocks (9) with the addition of Block H. *Sce* HO, the yeast mating type endonuclease, has been included in the table because of its similarity to inteins. The position in the protein of the last amino acid in each block is listed to the right of the block. The consensus line represents conserved residues or amino acid groups present in at least 15 of the 29 inteins included in the multiple sequence analysis. The four absolutely conserved residues are marked with an asterisk under the consensus line residue. Dashes indicate no match to that block. ∞The deposited DNA sequence yields a non-consensus *Psp* pol-3 intein Block E sequence of FLEGYSSAMA. However, if a frameshift resulting from insertion of a T at nt 3846 were made, the DNA sequence would then yield the conserved motif listed in this table, while three frameshifts in this region could give a sequence nearly identical to that of the other intein alleles. Intein names and abbreviations are as in Table 1. Definition of symbols in the consensus line: capital letters indicate conserved amino acids (standard single letter code); p, polar residue (S, T or C; purple); h, hydrophobic residue (G, A, V, L, I or M; green); a, acidic residue (D or E; red); b, basic residue (H, K or R; blue); r, aromatic residue (F, Y or W; orange).

Protein splicing requires four nucleophilic attacks mediated by three of the four conserved splice junction residues: (i) a Ser, Thr or Cys at the intein N-terminus; (ii) an Asn at the intein C-terminus; (iii) a Ser, Thr or Cys at the downstream extein N-terminus. The intein penultimate His assists in the C-terminal cleavage reaction.

Although Ser, Thr and Cys are chemically similar, it was initially speculated that splicing of thermostable inteins could not involve Cys because of high growth temperatures (24). It is now clear that inteins from thermophiles can utilize Cys, since all archaeal inteins listed in Table 1 are from thermophiles. However, with the still small sample size currently available, Thr has yet to be observed at an intein N-terminus and Ser has yet to be observed at the N-terminus of an intein from a mesophile (Table 1).

The requirement of a conserved His at the C-terminal splice junction must now be modified in light of the *Ceu* clpP, *Mja* PEP Syn and *Mja* Rpol A′ inteins that have Gly or Phe at this position (Table 1). However, splicing of these inteins has yet to be demonstrated in their native organisms, although splicing of the *Ceu* clpP intein in *Escherichia coli* requires changing the intein penultimate Gly to His (Liu,P.X.-Q., personal communication). Although Phe and Gly residues are unlikely to fulfil the role of assisting in C-terminal cleavage, since they cannot assist in acid/base catalysis, there is no *a priori* chemical requirement for this residue to be adjacent to the Asn in the primary amino acid sequence; the residue performing this function merely has to be near the Asn in three-dimensional space.

CONSERVED INTEIN MOTIFS

Twenty six new intein sequences have been determined since Pietrokovski first defined the seven conserved intein motifs termed Blocks A–G (9). The majority of the inteins included in the present analysis are found in archaea. Whether or not this biases the motifs will have to await the discovery of new eubacterial and eucaryal inteins. Seven highly similar allelic inteins were not included in the initial motif analysis using MACAW, but are listed in lower case in Table 2. The intein blocks depicted in Table 2 yielded the maximum score obtainable using the MACAW program. However, all of these motifs could be

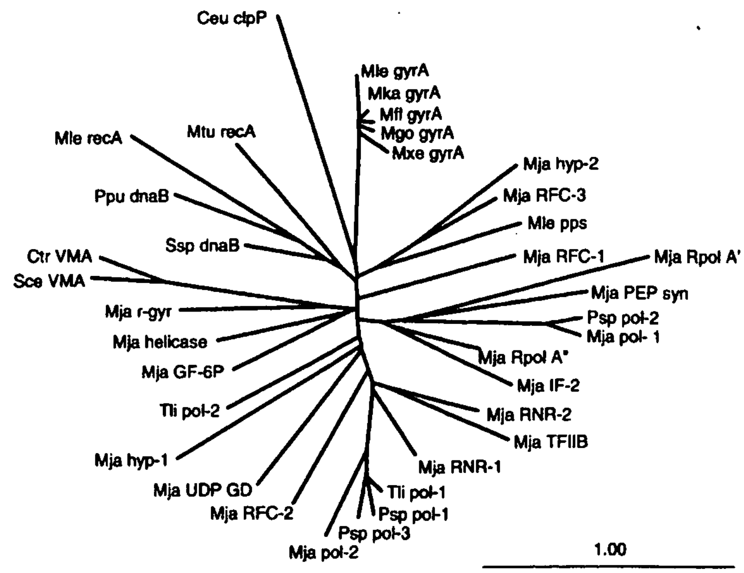

Figure 1. Unrooted phylogenetic tree based on the conserved intein motifs. The 95 columns of aligned residues in Table 2 were subjected to phylogenetic analysis using a least squares distance method (see Analytical Methods). Branch lengths shown are proportional to the estimated number of amino acid replacements per sequence position; the scale bar corresponds to an average of one replacement per position. Except for the grouping of alleles and the grouping of *Mja* Hyp-2 with *Mja* RFC-3 and *Mja* TFIIB with *Mja* RNR-2, all branches appear in <50% of the bootstrap replicates. Abbreviations as in Table 1.

expanded (except for limitations due to adjacent motifs or the sequence boundaries) and still yield highly significant scores. The size of some of the previously described motifs (9) has been modified in our analysis. For example, Block A has been reduced to 13 amino acids, although there is a less conserved, but still highly significant, block extending to residue 23.

Most positions in the intein motifs contain functionally or structurally similar amino acids, rather than a single predominant residue. In fact, only one His in Block B, two Gly in Block C (excluding inteins lacking this block) and one Asn in Block G are present in all inteins (marked by an asterisk under the consensus residue in Table 2). The consensus line in Table 2 lists amino acid groups (acidic, basic, aromatic, hydrophobic and polar) and conserved residues that are present in at least 15 of the 29 inteins used in the MACAW analysis. Note that many of these conserved residues can participate or assist in nucleophilic catalysis and the conserved Pro and Gly residues can affect secondary structure, being potential helix breaking residues. All blocks contain several hydrophobic residues.

Block A begins at the N-terminus of the intein and contains the chemically essential Ser or Cys residue. The sequence following the autocleavage site in hedgehog proteins fits the Block A consensus (50). Block B contains a polar residue (most often Thr) three amino acids prior to the only His conserved in all inteins. A similar motif is present in serine proteases and hedgehog proteins (51). The mechanism of cleavage in hedgehog proteins and at the intein N-terminus is similar (3,46,48,52). Thus, it is reasonable to suspect, and has been previously suggested (9), that the His in Block B may be involved in N-terminal splice junction cleavage. Block D is characterized by a conserved basic amino acid (most often Lys) and a Pro residue.

A 19 amino acid motif, called Block H, was found between blocks E and F. It overlaps with a previously identified, but unpublished, motif reported in the PRINTS database (http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html) (53). Block H is characterized by one or more Ser or Thr residues in positions 1–3, a central hydrophobic region containing several Leu and a Gly at position 18 followed by a hydrophobic residue. Block F contains an aromatic residue on both sides of several acidic and hydrophobic residues. If gaps were introduced into Block F, the presence of the extended consensus sequence, rVYDLpVa(1–3 residues)(H or E)NFh (see Table 2 legend for abbreviations) would be clearer. Block G is characterized by the three conserved C-terminal splice junction residues preceded by four hydrophobic residues and contains the first extein residue following the intein.

Blocks C and E are the dodecapeptide motifs that are required for endonuclease activity (8,54,55). Note that eight inteins have not maintained both blocks or the conserved acidic residues in these blocks which have been implicated in endonuclease activity (8,54,55), suggesting that these inteins may no longer be active endonucleases. A different *Mle* recA intein Block E sequence was assigned by the MACAW program in the present analysis. The previously published sequence of Block E was VLAIWYMDDG (9,23). Although this new motif assignment does not maintain the ~100 residue distance between Blocks C and E present in other inteins, it provides an equally good match to consensus dodecapeptide motifs (8). The absence of Block D could account for the reduced distance between blocks C and E in this intein.

The *S.cerevisiae* HO endonuclease contains all of the intein motifs except the conserved splice junction residues (Table 2; 9). HO endonuclease, which is essential for mating type switching,

is also a member of the dodecapeptide endonuclease family. Despite the presence of these conserved motifs and after addition of the conserved splice junction residues from the *Psp* pol-1 intein or the *Sce* VMA intein, HO does not splice at the protein level when placed in-frame between the *E.coli* maltose binding protein and a fragment of *Dirofilaria immitis* paramyosin (Platko,J. and Perler,F.B., unpublished data).

PHYLOGENY OF INTEIN SEQUENCES

Pairwise comparison of most inteins indicated a low degree of sequence similarity. Multiple sequence analysis identified motifs composed of groups of conserved residues, but not highly conserved specific amino acids. These factors made it difficult to determine the relationships among inteins present in the same or related organisms, in different domains of life or in different extein homologs. Therefore, the phylogenetic relationships of the 36 inteins were determined using programs in the PHYLIP package (29). This analysis revealed that, except for the intein alleles, there is no clustering of inteins on the basis of phylogenetic domain, organism classification, genus, species or location in the extein gene (Fig. 1). It further suggests that the 18 *M.jannaschii* inteins did not arise from recent intein duplications. Among non-allelic inteins, the only branches which appear in >50% of the bootstrap replicates are those associating *Mja* Hyp-2 with *Mja* RFC-3 (which was seen 83% of the time) and *Mja* TFIIB with *Mja* RNR-2 (54%). However, the observed relationships are not chaotic. Except for the dnaB inteins, all sets of allelic inteins grouped together in 99–100% of the 100 bootstrap samples.

Allelic inteins are more closely related than non-allelic inteins. Is this due to recent intein mobility events or to the acquisition of an intein by a common ancestor? Since intein alleles are not present in all closely related isolates or organisms, there must be a mechanism for intein gain or loss. For example, inteins are absent in DNA polymerases from 11 of 17 Archaea analyzed, with only six alleles of *Tli* pol-1, one allele of *Tli* pol-2 and two alleles of *Psp* pol-2 (17,35). Depending on the *Mycobacterium* species, not all isolates contain the recA or gyrA inteins (14,15) and of six Archaea examined, only *M.jannaschii* contains an RNA polymerase intein (17,41–45).

Gain of inteins is supported by several lines of evidence. Intein mobility has been demonstrated in yeast (6). Intein gene mobility is initiated when an inteinless allele enters the cell via sexual reproduction, conjugation, transduction, phage infection, plasmid transfer, etc. The inteinless allele is then cleaved by the intein endonuclease (homing endonucleases do not cut their own genomic DNA when the intein is present) (6–8,12,32). This endonuclease activity, combined with extein homology, substantially increases the rate of gene conversion by the double-strand break repair recombination pathway (6,11,12,38,56,57). As predicted, allelic inteins *Tli* pol-1 and *Psp* pol-1 are isoschizomers with the same endonuclease specificity (Perler,F.B., unpublished data). A second line of evidence for lateral transmission of inteins is the observation that codon usage in the gyrA inteins is different from extein codon usage, suggesting that the inteins have been recently acquired from a different species (15). Finally, the DNA polymerases from GB-D and GI-J Thermococcales isolates (98% identical over the 96 amino acid GI-J fragment sequenced) are more similar than the GB-D and *T.litoralis* DNA polymerases (78% identical), although there is no intein in the GI-J DNA polymerase while there are allelic inteins in the GB-D and *T.litoralis* DNA polymerases (60% identity between inteins) (35).

If intein alleles are ancient and can be lost with time, the mechanism for intein loss has to be very specific to avoid inactivating mutations in the extein gene. Recombination could lead to intein loss if the intein was no longer an active homing endonuclease, however, if the intein was an active homing endonuclease, lateral transmission should predominate. Recombination in haploid organisms such as *Mycobacterium* spp., *M.jannaschii* and Thermococcales can only occur if merodiploids are occasionally formed. Yet the presence of inteins in haploid individuals is very variable. Barring an unknown efficient mechanism for intein loss other than by rare recombination events, the prevalence of intein loss would require selection against inteins.

Taken together, these data suggest that the presence of intein alleles is most often due to lateral transmission rather than the early acquisition of an intein by a common ancestor. On the other hand, there is no phylogenetic evidence that non-allelic inteins have spread by lateral transmission, although it is possible that they arose by an illegitimate lateral transmission event within the same genome followed by significant divergence.

IDENTIFYING INTEINS

How one identifies new inteins depends on whether you are analyzing the sequence of a specific gene or searching databases for new inteins. A large in-frame insertion in a sequenced gene that is absent in other sequenced homologs suggests that this gene may contain an intein. The sequence should then be examined for the presence of the conserved intein junction residues and the intein blocks, including the dodecapeptide motifs. Not all inteins will have a His as the penultimate residue. However, since most inteins end in His–Asn, the His–Asn–(Ser, Thr, Cys) C-terminal intein motif is still a valid tool for identifying intein boundaries. Not all intein blocks need be present (Table 2). Since several amino acids are found within each position in a block, the putative intein sequence should be checked for the presence of a member of the amino acid group present at that position (Table 2).

In examining databases, inteins can be identified by searching with the conserved intein blocks (9,18) or complete intein amino acid sequences. Once a match has been found, the entire sequence should be re-analyzed for the presence of other conserved intein motifs and database searches should be performed to find matches to the putative extein sequences. The presence of the conserved splice junction residues and the conserved blocks are not sufficient to label a sequence an intein in the absence of comparison with an inteinless extein homolog, although the presence of all the blocks would be highly indicative of the presence of an intein in an extein gene that does not have a sequenced homolog. In the absence of experimentally demonstrating protein splicing, it should be emphasized that the combined use of these criteria, rather than the use of any single criterion, yields the most significant results.

ACKNOWLEDGEMENTS

We thank Shmuel Pietrokovski, Stewart Cole, Paul X.-Q.Liu, Amalio Telenti and Mike Reith for helpful discussions and providing us with their unpublished results and Sanjay Kumar, Bill Jack, Chris Noren, Maurice Southworth and Ming Xu for

helpful discussions. We thank Shmuel Pietrokovski for sharing information concerning new inteins. We thank Donald G.Comb for support and encouragement.

REFERENCES

1 Perler,F.B., Davis,E.O., Dean,G.E., Gimble,F.S., Jack,W.E., Neff,N., Noren,C.J., Thorner,J. and Belfort,M. (1994) *Nucleic Acids Res.*, **22**, 1125–1127.
2 Guan,C., Cui,T., Rao,V., Liao,W., Benner,J., Lin,C.L. and Comb,D. (1996) *J. Biol. Chem.*, **271**, 1732–1737.
3 Porter,J.A., Ekker,S.C., Park,W.J., von Kessler,D.P., Young,K.E., Chen,C.H., Ma,Y., Woods,A.S., Cotter,R.J., Koonin,E.V. and Beachy,P.A. (1996) *Cell*, **86**, 21–34.
4 Hirata,R., Ohsumi,Y., Nakano,A., Kawasaki,H., Suzuki,K. and Anraku,Y. (1990) *J. Biol. Chem.*, **265**, 6726–6733.
5 Kane,P.M., Yamashiro,C.T., Wolczyk,D.F., Neff,N., Goebl,M. and Stevens,T.H. (1990) *Science*, **250**, 651–657.
6 Gimble,F.S. and Thorner,J. (1992) *Nature*, **357**, 301–306.
7 Bremer,M., Gimble,F.S., Thorner,J. and Smith,C. (1992) *Nucleic Acids Res.*, **20**, 5484.
8 Mueller,J.E., Bryk,M., Loizos,N. and Belfort,M. (1994) In Linn,S.M., Lloyd,R.S. and Roberts,R.J. (eds), *Nucleases*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 111–143.
9 Pietrokovski,S. (1994) *Protein Sci.*, **3**, 2340–2350.
10 Perler,F.B., Comb,D.G., Jack,W.E., Moran,L.S., Qiang,B., Kucera,R.B., Benner,J., Slatko,B.E., Nwankwo,D.O., Hempstead,S.K., Carlow,C.K.S. and Jannasch,H. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 5577–5581.
11 Belfort,M. and Perlman,P.S. (1995) *J. Biol. Chem.*, **270**, 30237–30240.
12 Lambowitz,A.M. and Belfort,M. (1993) *Annu. Rev. Biochem.*, **62**, 587–622.
13 Davis,E.O., Jenner,P.J., Brooks,P.C., Colston,M.J. and Sedgwick,S.G. (1992) *Cell*, **71**, 201–210.
14 Davis,E.O., Thangaraj,J.S., Brooks,P.C. and Colston,M.J. (1994) *EMBO J.*, **13**, 699–703.
15 Fsihi,H., Vincent,V. and Cole,S.T. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 3410–3415.
16 Gu,H.H., Xu,J., Gallagher,M. and Dean,G.E. (1993) *J. Biol. Chem.*, **268**, 7372–7381.
17 Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D., Kerlavage,A.R., Dougherty,B.A., Tomb,J., Adams,M.D., Reich,C.I., Overbeek,R., Kirkness,E.F., Weinstock,K.G., Merrick,J.M., Glodek,A., Scott,J.L., Geoghagen,N.S.M., Weidman,J.F., Fuhrmann,J.L., Nguyen,D., Utterback,T.R., Kelley,J.M., Peterson,J.D., Sadow,P.W., Hanna,M.C., Cotton,M.D., Roberts,K.M., Hurst,M.A., Kaine,B.P., Borodovsky,K.H., Fraser,C.M., Smith,H.O., Woese,C.R. and Venter,J.C. (1996) *Science*, **273**, 1058–1073.
18 Pietrokovski,S. (1996) *Trends Genet.*, **12**, 287–288.
19 Reith,M.E. and Munholland,J. (1995) *Plant Mol. Biol. Rep.*, **13**, 333–335.
20 Huang,C., Wang,S., Chen,L., Lemieux,C., Otis,C., Turmel,M. and Liu,X.Q. (1994) *Mol. Gen. Genet.*, **244**, 151–159.
21 Xu,M., Comb,D.G., Paulus,H., Noren,C.J., Shao,Y. and Perler,F.B. (1994) *EMBO J.*, **13**, 5517–22.
22 Anraku,Y. and Hirata,R. (1994) *J. Biochem.*, **115**, 175–178.
23 Davis,E.O. and Jenner,P.J. (1995) *Antonie Van Leeuwenhoek*, **67**, 131–137.
24 Cooper,A.A., Chen,Y., Lindorfer,M.A. and Stevens,T.H. (1993) *EMBO J.*, **12**, 2575–2583.
25 Genetics Computer Group (1994) *Program Manual for the Wisconsin Package*, Version 8. Genetics Computer Group, Madison, WI.
26 Schuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) *Protein Struct. Funct. Genet.*, **9**, 180–90.
27 Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) *Science*, **262**, 208–14.
28 Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 10915–9.
29 Felsenstein,J. (1989) *Cladistics*, **5**, 164–166.
30 Swofford,D.L., Olsen,G.J., Waddell,P.J. and Hillis,D.M. (1996) In Hillis,D.M., Moritz,C. and Mable,B.K. (eds), *Molecular Systematics*, 2nd Edn. Sinauer Associates, Sunderland, MA, pp. 407–514.
31 Roberts,R.J. and Macelis,D. (1996) *Nucleic Acids Res.*, **24**, 223–235.
32 Gimble,F.S. and Thorner,J. (1993) *J. Biol. Chem.*, **268**, 21844–21853.
33 Xu,M., Southworth,M.W., Mersha,F.B., Hornstra,L.J. and Perler,F.B. (1993) *Cell*, **75**, 1371–1377.
34 Davis,E.O., Sedgwick,S.G. and Colston,M.J. (1991) *J. Bacteriol.*, **173**, 5653–5662.
35 Perler,F.B., Kumar,S. and Kong,H. (1996) In Adams,M.W.W. (ed.), *Enzymes and Proteins from Hyperthermophilic Microorganisms.* Academic Press, New York, NY, Vol. 48, pp. 377–435.
36 Sun,D. and Setlow,P. (1993) *J. Bacteriol.*, **175**, 2501–2506.
37 Belfort,M., Reaban,M.E., Coetzee,T. and Dalgaard,J.Z. (1995) *J. Bacteriol.*, **177**, 3897–3903.
38 Kaneko,T., Tanaka,A., Sato,S., Kotani,H., Sazuka,T., Miyajima,N., Sugiura,M. and Tabata,S. (1995) *DNA Res.*, **2**, 191–198.
39 Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M., McKenney,K., Sutton,G., FitzHugh,W., Fields,C., Gocayne,J.D., Scott,J., Shirley,R., Liu,L.-I., Glodek,A., Kelley,J.M., Weidman,J.F., Phillips,C.A., Spriggs,T., Hedblom,E., Cotton,M.D., Utterback,T.R., Hanna,M.C., Nguyen,D.T., Saudek,D.M., Brandon,R.C., Fine,L.D., Fritchman,J.L., Fuhrmann,J.L., Geoghagen,N.S.M., Gnehm,C.L., McDonald,L.A., Small,K.V., Fraser,C.M., Smith,H.O. and Venter,J.C. (1995) *Science*, **269**, 496–512.
40 Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M., Fritchman,J.L., Weidman,J.F., Small,K.V., Sandusky,M., Fuhrmann,J., Nguyen,D., Utterback,T.R., Saudek,D.M., Phillips,C.A., Merrick,J.M., Tomb,J.-F., Dougherty,B.A., Bott,K.F., Hu,P.-C., Lucier,T.S., Peterson,S.N., Smith,H.O., Hutchison,C.A. and Venter,J.C. (1995) *Science*, **270**, 397–403.
41 Pühler,G., Lottspeich,F. and Zillig,W. (1989) *Nucleic Acids Res.*, **17**, 4517–34.
42 Klenk,H.-P., Schwass,V., Lottspeich,F. and Zillig,W. (1992) *Nucleic Acids Res.*, **20**, 4659.
43 Klenk,H.-P., Renner,O., Schwass,V. and Zillig,W. (1992) *Nucleic Acids Res.*, **20**, 5226.
44 Leffers,H., Gropp,F., Lottspeich,F., Zillig,W. and Garrett,R.A. (1989) *J. Mol. Biol.*, **206**, 1–17.
45 Berghofer,B., Krockel,L., Kortner,C., Truss,M., Schallenberg,J. and Klein,A. (1988) *Nucleic Acids Res.*, **16**, 8113–8128.
46 Xu,M. and Perler,F.B. (1996) *EMBO J.*, **15**, 5146–5153.
47 Shao,Y., Xu,M.Q. and Paulus,H. (1995) *Biochemistry*, **34**, 10844–10850.
48 Shao,Y., Xu,M.-Q. and Paulus,H. (1996) *Biochemistry*, **35**, 3810–3815.
49 Chong,S., Shao,Y., Paulus,H., Benner,J., Perler,F.B. and Xu,M. (1996) *J. Biol. Chem.*, **271**, 22159–22168.
50 Koonin,E.V. (1995) *Trends Biochem Sci*, **20**, 141–142.
51 Lee,J.J., Ekker,S.C., von Kessler,D.P., Porter,J.A., Sun,B.I. and Beachy,P. A. (1994) *Science*, **266**, 1528–1537.
52 Porter,J.A., von Kessler,D.P., Ekker,S.C., Young,K.E., Lee,J.J., Moses,K. and Beachy,P.A. (1995) *Nature*, **374**, 363–366.
53 Attwood,T.K., Beck,M.E., Bleasby,A.J., Degtyarenko,K. and Parry Smith,D.J. (1996) *Nucleic Acids Res.*, **24**, 182–188.
54 Hodges,R.A., Perler,F.B., Noren,C.J. and Jack,W.E. (1992) *Nucleic Acids Res.*, **20**, 6153–6157.
55 Gimble,F.S. and Stephens,B.W. (1995) *J. Biol. Chem.*, **270**, 5849–5856.
56 Quirk,S.M., Bell-Pedersen,D. and Belfort,M. (1989) *Cell*, **56**, 455–65.
57 Bell-Pedersen,D., Quirk,S.M., Aubrey,M. and Belfort,M. (1989) *Gene*, **82**, 119–26.